DOCUMENT RESUME

ED 448 196 TM 032 226

AUTHOR Buckendahl, Chad W.; Smith, Russ W.; Impara, James C.; Plake, Barbara S.

TITLE A Comparison of the Angoff and Bookmark Standard Setting Methods.

PUB DATE 2000-10-00

NOTE 13p.; Paper presented at the Annual Meeting of the Mid-Western Educational Research Association (Chicago, IL, October 25-28, 2000).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Comparative Analysis; *Cutting Scores; *Junior High School Students; Junior High Schools; Mathematics Tests; *Research Methodology; *Standard Setting (Scoring); *Standards

IDENTIFIERS *Angoff Methods

ABSTRACT

This paper presents a comparison of two commonly used methods, Angoff (W. Angoff, 1971) and Bookmark (D. Lewis, H. Mitzel, and D. Green, 1996), for setting cut scores on selected response tests. This comparison is presented through an application to a grade 7 mathematics assessment in a suburban Midwestern school district. Training and operational methods and procedures for each method are described in detail along with comparative results for the application. Although the Angoff method is more widely used, the Bookmark method has several strengths. It provides judges with an opportunity to focus on performance of the "Barely Proficient" students without worrying about estimating item difficulty. It also may be a more efficient method in terms of the length of time it takes for judges to make their bookmark placements. When results were compared, the recommended cut scores provided by the two methods were very similar, with the Bookmark method producing a lower standard deviation. (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 448 196

A comparison of the Angoff and Bookmark standard setting methods

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

C W Buckendahl

1

Chad W. Buckendahl

Russ W. Smith

James C. Impara

Barbara S. Plake

University of Nebraska – Lincoln

Paper presented at the annual meeting of the

Mid-Western Educational Research Association in Chicago, IL

BEST COPY AVAILABLE

TM032226

October, 2000

Abstract

This paper presents a comparison of two commonly used methods, Angoff and Bookmark, for setting cut scores on selected response tests. This comparison is presented through an application with a Grade 7 Mathematics Assessment in a suburban Mid-Western school district. Training and operational methods and procedures for each method are described in detail along with comparative results for the application.

Although the Angoff method is more widely used, the Bookmark method has several strengths. It provides judges with an opportunity to focus on performance of the "Barely Proficient" student without worrying about estimating item difficulty. It also may be a more efficient method in terms of the length of time it takes for judges to make their bookmark placements. In comparing results, the recommended cut scores provided by the two methods were very similar with the Bookmark method producing a lower standard deviation.

A comparison of Angoff and Bookmark standard setting methods

Test-centered methods for setting Minimum Passing Scores (MPSs), or cut scores, on selected response assessments have been well researched. The most prevalent method for setting cut scores on these assessments is the Angoff (1971) method. CTB/McGraw-Hill has recently developed an alternative standard setting method for setting cut scores on assessments with selected response, constructed response or a combination of the two item types. The Bookmark method (Lewis, Mitzel, & Green, 1996) uses an item response theory based item mapping procedure to order items and attempts to simplify the cognitive tasks required by standard setting judges. Since cut scores may be used to make high-stakes decisions about students including assignment to remedial, other "re-looping" educational programs, or even eligibility for graduation it is important that the methods provide sound evidence of cut score validity.

The purpose of this study was to compare recommended cut scores from a standard setting workshop for a suburban Mid-western school district that used both the Angoff and Bookmark methods.

Test Information

The Grade 7 Mathematics Examination is intended to be used to provide information on the extent that district seventh graders have attained the skills in mathematics consistent with the current curriculum emphasis in that subject. The cut score for this test was set to classify students into two categories: 1) students who need additional mathematics instruction (re-education) so that they can be brought to a point of being "on-track" in their performance of mathematics standards, and 2) students who are

considered to be "on-track" in mathematics. The content of the Mathematics test includes 69 multiple-choice items that assess six strands of mathematics skills. All items are scored dichotomously and each item counts one point. The total number of points available is 69.

Methods and Procedures

Angoff Method

The Angoff (1971) method entails using expert judges to examine each item on the test and estimate how a typical borderline "Barely Proficient" student (BPS) will perform on that item. For the Grade 7 Mathematics Examination teachers were asked (after a training activity) to conceptualize a specific barely proficient student they had taught. Keeping this student in mind, they were directed to indicate, for each item, whether the student they had in mind would answer the item correctly or not (Right or Wrong). This was done for each item the teachers rated. After an initial rating, actual performance data (proportion of students answering each item correctly and the cumulative percent of students at each score point) from a representative sample of over 400 of the district's students was provided to the teachers. After seeing the data, teachers were asked to make a second estimate of whether the barely proficient student would answer correctly or not. The second estimate could be either the same or different from their first estimate. The feedback data used for both methods provides a reality check to ensure that expected performance is not set either unrealistically high or low because the teacher has misjudged how hard or easy the item actually is. The recommended cut score is based on the second estimate. It is calculated by summing, for each teacher, the number of "Right" items and then averaging those values across the teachers.

Bookmark Method

The Bookmark method also uses expert judges to examine items on the test and estimate how a typical borderline "Barely Proficient" student will perform on that item. Items are ordered from least difficult to most difficult and compiled into a booklet. Item difficulties (p-values) were estimated from a representative sample of over 400 of the district's students used in the pilot test. After the same training activity as teachers in the Angoff group experienced, these teachers were asked to conceptualize a specific barely proficient student they had taught. Keeping this student in mind, they were directed to start with the easiest item and move through the booklet until they found the place where their barely proficient student would probably get all items to that point correct and probably get all items beyond that point incorrect[1]. At point in the booklet, the teachers placed their bookmarks. After the initial bookmark placement, the same actual performance data (proportion answering each item correctly and the cumulative percent of students at each score point) as was given to the Angoff group was provided to the teachers. After seeing the data, teachers were asked to make a second bookmark placement for how they expected the barely proficient student to perform. The second bookmark could be either the same or different from their initial bookmark. The cut score is based on the second bookmark placement. It is calculated by summing, for each teacher, the number of items up to the bookmark and then averaging those values across the teachers.

Training procedures

The workshop for setting cut scores used a panel of 23 teachers. These teachers

[1] CTB/McGraw Hill's version of the Bookmark Method asks judges to estimate the point at which the target student would get .67 of the items correct.

were selected by the district to represent a cross section of the district's teachers and their classes represent a cross section of the district's students. At the meeting a) the teachers were told the purpose of the meeting, b) the test specifications were reviewed, c) a process for helping the teachers conceptualize the "Barely Proficient" student was undertaken, d) specific training in the item performance estimation procedure (Angoff or Bookmark) was provided, e) teachers made item performance estimates, and f) teachers evaluated the standard setting workshop.

After the undergoing the same orientation, reviewing the table of specifications, and conceptualization of the "Barely Proficient" student, the group was split (12 teachers for Angoff and 11 teachers for Bookmark) and adjourned to separate rooms to undertake additional training for the Angoff and Bookmark methods. Teachers were assigned to groups to ensure a representative sample of schools in each group.

A practice activity was undertaken to familiarize judges with the type of items and the range of difficulty they would see on the operational test. Practice items were taken from the 7th grade version of a similar mathematics test and presented in order of administration. In the Angoff group, for each item, panelists indicated a "Right" or a "Wrong" (R or W) for the specific BPS they had in mind. An "R" suggested the panelist believed the student would answer the item correctly and a "W" indicated the panelist believed the student would answer incorrectly. When panelists had completed their performance estimates for the items on this practice test, each item was discussed. The discussion revolved around the panelists' reasons for indicating R or W. Panelists were asked to explain why they had responded R or W in the context of the general characteristics elicited in the earlier discussion of the BPS. Panelists were told that

variability among the teachers was expected, that BPSs were not expected to all be the same in their ability to answer questions, so some may be able to respond correctly and others not for a particular item.

In the practice activity for the Bookmark group, items were rank-ordered from easiest to most difficult and compiled in a booklet. These were the same items used for practice in the Angoff method. The only difference was that for the Angoff method, the practice items were presented in order of administration whereas the Bookmark method ordered items from easiest to most difficult by p-value (proportion of students in the pilot sample that answered the item correctly). For the specific BPS they had in mind, they started with the easiest item and progressed through the booklet until they reached the point at which they believed their BPS would probably get all items up to that point correct and all items beyond that point incorrect. At that point they placed their bookmark. When panelists had completed their bookmark placement for the items on this practice test, each item was discussed. The discussion revolved around the panelists' reasons for where they placed their bookmark. Panelists were asked to explain why they had placed their bookmark in a specific place in the context of the general characteristics of the BPS. Panelists were told that variability among the teachers was expected, that BPSs were not expected to all be the same in their ability to answer questions, so some may be able to respond correctly and others not for a particular item.

The panelists were then provided with actual performance data on each item. The performance data consisted of the proportion of the MPS students who had answered each item correctly (called p-values). The practice test consisted of items that had a range in difficulty similar to the range found in the operational test. After discussion of

all practice items, the teachers were shown the impact of a range of cut scores. The impact data were based on cumulative percents that were derived from the sample of MPS students' performance on these items.

Angoff Procedures

After all teachers made their initial estimates and these forms were collected, copies of the test and a separate answer key were distributed, and teachers made their first round ratings. As teachers made their ratings, their rating forms were collected and the ratings entered into a computer program designed to compute the cut score. After teachers completed their first round of ratings, their rating forms were returned and actual performance data provided and explained. The actual performance data included item p-values and a cumulative percent distribution of actual student performance. The impact (percent classified as Below Proficient) of the teachers' collective first round estimate was shown. Teachers then made their second (final) rating of the 69 items.

Bookmark Procedures

After all teachers made their initial estimates and these forms were collected, copies of the test booklets (with items rank ordered by difficulty) and a separate answer key were distributed, and teachers made their first round bookmark placement. After teachers made this placement their forms were collected and the rating entered into a computer program designed to compute the cut score. Following data entry, their rating forms were returned and actual performance data provided and explained. The actual performance data included item p-values and a cumulative percent distribution of actual student performance. The impact (percent classified as Below Proficient) of the teachers' collective first round estimate was shown. Teachers then made their second (final)

bookmark placement on the 69-item test.

Results

Angoff Method

The teachers provided performance estimates before and after being given actual performance data. The cut scores from each round are shown in Table 1 below. Providing actual performance data between rounds one and two had some influence on the teachers as the second round cut score dropped by a point and a half. The variation in cut scores changed from Round 1 to Round 2, increased from a standard deviation of 7.79 in Round 1 to 10.96 in Round 2. This change in variance is not unusual because some teachers will use the performance data to adjust their judgments higher or lower than their first round cut score.

Table 1.

Angoff method cut score means and standard deviations for Rounds 1 and 2.

Round	Cut score	Standard Deviation	% below
1	34.92	7.79	8.9%
2	33.42	10.96	7.6%

Bookmark Method

The teachers provided bookmark placements before and after being given actual performance data. The cut scores from each round are shown in Table 2 below. Providing actual performance data between rounds one and two had some influence on the teachers as the second round cut score increased by two points. The variation in cut scores also changed from Round 1 to Round 2, decreased from a standard deviation of

11.03 in Round 1 to 8.66 in Round 2. This small change in variance was also expected for this method because it is easier for panelists to see how the placement of their bookmark affects the final cut score. In the Bookmark Method, panelists know that by moving their bookmark ahead, it will increase the cut score and more students would not pass the exam. The alternative scenario is also true. This is different from the Angoff method in which panelists re-visit every item on the test and make a judgment as to whether or not they are comfortable with their initial judgment.

Table 2.

Bookmark method cut score means and standard deviations for Rounds 1 and 2.

Round	Cut score	Standard Deviation	% below
1	33.64	11.03	7.6%
2	35.64	8.66	9.4%

Although there was a small difference (2 points, which impacted 1.8% of students in the sample) in final mean cut scores between the Angoff and Bookmark groups, the standard deviation was lower for the second round of the Bookmark method compared with the Angoff method. This smaller standard deviation would produce a smaller range of possible cut scores and indicate a higher level of inter-judge agreement when considered by a policymaking body. Workshop evaluation data was also collected to determine the level of confidence and comfort of members of each method's group. Results indicated similar levels of confidence in the passing score and comfort in the process between the groups.

Conclusions and Implications

This paper presents a comparison of two commonly used methods for setting cut scores on selected response tests through an application with a Grade 7 Mathematics Assessment in a suburban Mid-western school district. Although the Angoff method is more widely used, the modification of the Bookmark method has several strengths. It provides judges with an opportunity to focus on performance of the "Barely Proficient" student without worrying about estimating item difficulty. Although not systematically examined in this study, it may be a more efficient method in terms of the length of time it takes for judges to make their bookmark placements. In comparing results, the recommended cut scores provided by the two methods were very similar with the Bookmark method producing a lower standard deviation than the Angoff method. The application of the Bookmark standard setting method as it is related to the Angoff method warrants further study. As stated above, many schools are developing or using assessments for high stakes decisions (e.g., remediation, grade promotion, or graduation). If the Bookmark method is more efficient and reduces the item difficulty estimation responsibilities of judges, it may be better served in specific settings as compared to others. Because the illustration of the application in this paper was shown for a lower stakes assessment, studies examining higher stakes assessments would also be beneficial.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement, 2nd Edition, Washington, DC: American Council on Education.

Lewis, D.M., Mitzel, H.C., & Green, D.R. (1996). Standard setting: A Bookmark approach. Symposium presented at the Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix, AZ.